TITLE:

ラベル選択付最小連結全域部分グラフ問題と化学構造式OCRへの応用 (アルゴリズムと計算理論の新展開)

AUTHOR(S):

藤芳, 明生; 鈴木, 昌和

CITATION:

藤芳, 明生 ...[et al]. ラベル選択付最小連結全域部分グラフ問題と化学構造式OCRへの応用 (アルゴリズムと計算理論の新展開). 数理解析研究所講究録 2012, 1799: 111-117

ISSUE DATE:

2012-06

URL:

http://hdl.handle.net/2433/172993

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

数理解析研究所講究録
第1799巻 2012年 111-117

ラベル選択付最小連結全域部分グラフ問題と化学構造式OCRへの応用

藤芳 明生
茨城大学工学部情報工学科
fujiyosi@mx.ibaraki.ac.jp

鈴木 昌和
九州大学大学院数理学研究院
msuzuki@kyudai.jp

概 要

本稿では，ラベル選択付最小全域木問題を拡張したラベル選択付最小連結全域部分グラフ問題を考える．最小連結全域部分グラフ問題とは，各辺に「繋ぐ場合の重み」と「繋がない場合の重み」の2種類を与え，選ばれた辺の「繋ぐ場合の重み」の合計と選ばれなかった辺の「繋がない場合の重み」の合計の和が最小となるような連結全域部分グラフを求める問題である．ラベル選択付最小全域木問題がNP困難であるため，この問題を拡張したラベル選択付最小連結全域部分グラフ問題も同様にNP困難である．しかし，化学構造式OCRの認識精度向上に応用するためには，この問題を現実的な時間内に解く必要がある．そこで，入力のグラフのtree-widthを2以下に制限した場合を考え，線形時間で動作する非常にシンプルなアルゴリズムを提案する．

1 はじめに

日本では，諸外国と同じように，特許申請書は出願から1年6ヶ月後に自動的に公開される．2008年には，312,443件の特許申請書が公開された．これらの公開された特許申請書は，特許庁からDVD－ROMで購入することも可能であり，また，特許電子図書館 [1] のWebページから自由に検索，ダウンロードすることも可能である．特許申請書のほとんど部分はXMLフォーマットの文字列であるが，化学構造式，数式，図，表はTIFF画像として収録されている．2008年に公開された特許申請書を調べたところ，化学構造式は，229,969枚のTIFF画像として収録されていた．化学・製薬産業において，特許中の化学構造式の包括的な検索は必須である．そのため，特許中の化学構造式を含む画像を，すばやく低コストで電子化する化学構造式OCRの開発が求められている．本稿では，化学構造式OCR[2, 3]の認識精度向上に応用するため，ラベル選択付最小連結全域部分グラフ問題を考える．

ラベル選択付最小連結全域部分グラフ問題は，ラベル選択付最小全域木問題 [4, 5] を拡張することで得られる．この問題の例を図1 (a) に示す．ラベル記号の集合は，$\Sigma = \{a, b, c, d\}$ である．頂点は点線の長方形で示されており，それぞれの頂点は少なくとも1つ以上のラベル候補を持っている．ラベル候補はラベル記号を円で囲って示されている．重み付き辺はラベル候補どうしを結んでおり，各辺には左側に記された「繋ぐ場合の重み」と右側の括弧の中に記された「繋がない場合の重み」の2種類の重みが与えられている．2つの頂点のラベル候補のいずれかの間に重み付き辺が存在するならば，その頂点間のすべてのラベル候補の間には重み付き辺が存在するものとする．例では，いくつかのラベル候補の組み合わせには

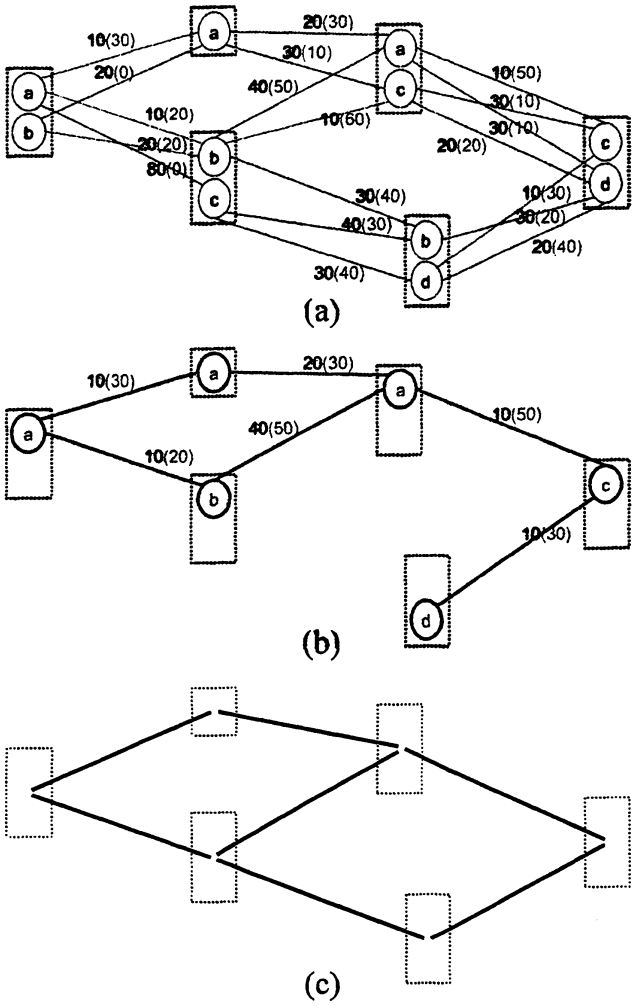

図 1: (a) ラベル選択付最小連結全域部分グラフ問題の例, (b) ラベル選択付最小連結全域部分グラフ, (c) 基礎グラフ.

辺が記されていないが，そこには,「繋ぐ場合の重み」が無限大で「繋がない場合の重み」が 0 であるような辺が省略されているものと考える．この問題に対するラベル選択付最小連結全域部分グラフを図 1 (b) に示す．それぞれの頂点から，正確に一つずつラベル候補が選択されており，選ばれた辺によって連結全域部分グラフが構成されている．選ばれた辺の「繋ぐ場合の重み」の合計と選ばれなかった辺の「繋がない場合の重み」の合計は，すべてのラベル選択と連結全域部分グラフの組み合わせの中で最小となっている．更に，図 1 (c) に示すような，基礎グラフという概念も導入する． 基礎グラフとは，ラベル候補間に重み付き辺が存在する頂点間を辺で結ぶことでできるグラフである．

ラベル選択付最小全域木問題が NP 困難であるため，この問題を拡張したラベル選択付最小連結全域部分グラフ問題も同様に NP 困難である．ラベル選択付最小連結全域部分グラフ問題において，繋がない場合の重みがすべて 0 ならば，なるべく繋がない方が良いので，ラベル選択付最小連結全域部分グラフはサイクルを持つことはない．つまり，ラベル選択付最小全域木問題を解いたことになる．

化学構造式 OCR の認識精度向上に応用するためには，この問題を現実的な時間内に解

表 1: ChEMBL 中の化合物の tree-width.

Tree-Width	数	割合
1	11,753	(1.85%)
2	603,806	(94.95%)
3	20,031	(3.15%)
≥ 4	343	(0.05%)
Total	635,933	(100.00%)

く必要がある．そこで，基礎グラフの tree-width を 2 以下に制限した場合を考え，線形時間で動作する非常にシンプルなアルゴリズムを提案する．本稿が提案するアルゴリズムは，tree-width が 2 以下のグラフにしか対応できないが，tree-width が 3 以上の場合であっても，一部の辺の処理を後回しにし，tree-width が 2 以下の部分グラフに対してアルゴリズムを適応すれば，化学構造式 OCR への応用には十分であると考える．医薬品及び医薬品候補化合物に限定すれば，それらグラフ表現の tree-width は十分に小さいことが知られている．表 1 に，ChEMBL[6] に登録されている 635,933 個の医薬品及び医薬品候補化合物の tree-width を示す．ほとんどの化合物の tree-width は 2 ないし 3 以下となっていることが確認された．

2 諸定義

本章では，本論文で利用される記法，定義を与えるとともに，問題の形式的な再定義を行う．

グラフとは，順序対 $G = (V, E)$ のことである．ここで，V は**頂点**の有限集合，異なる頂点の組を**辺**と呼び，E は辺の有限集合である．グラフが**連結**であるとは，グラフ上の任意の 2 頂点間に道が存在することである．グラフ $G' = (V', E')$ がグラフ G の**全域部分グラフ**であるとは，$V' = V$ かつ $E' \subseteq E$ となることである．連結であるような全域部分グラフを**連結全域部分グラフ**と呼ぶ．

Σ を**ラベル**の有限集合とする．本稿では，$|\Sigma|$ は最初から決められた定数であると考える．R_+ を正実数全体の集合とする．グラフ $G = (V, E)$ の**重み関数**とは，部分関数 $w : V \times V \times \Sigma \times \Sigma \to R_+ \cup \{0, \infty\}$ のことであり，すべての $v_1, v_2 \in V$ と $l_1, l_2 \in \Sigma$ について，$\{v_1, v_2\} \in E$ ならば $w(v_1, v_2, l_1, l_2) = w(v_2, v_1, l_2, l_1)$ となり，$\{v_1, v_2\} \notin E$ ならば $w(v_1, v_2, l_1, l_2)$ は未定義となる．G の**ラベル割り当て**とは，関数 $\sigma : V \to \Sigma$ のことである．

定義 1 $G = (V, E)$ をグラフ，w と $\overline{w}$ を G の重み関数とする．w を**繋ぐ場合の重み関数**，$\overline{w}$ を**繋がない場合の重み関数**と呼ぶ．G のラベル割り当て σ に対する連結全域部分グラフ $G' = (V, E')$ の**重さ**を次で定義する．

$$w(G') = \sum_{\{v_1, v_2\} \in E'} w(v_1, v_2, \sigma(v_1), \sigma(v_2)) + \sum_{\{v_1, v_2\} \in E - E'} \overline{w}(v_1, v_2, \sigma(v_1), \sigma(v_2)).$$

定義 2 $G=(V,E)$ をグラフ，w と $\overline{w}$ を G の重み関数とする．**ラベル選択付最小連結全域部分グラフ問題**とは，連結全域部分グラフ $G'=(V,E')$ と G のラベル割り当て σ の組み合わせ中で，G' の重さが最小となるものを見つける問題である．

3 NP困難性

ラベル選択付最小連結全域部分グラフ問題のNP困難性は，既にNP困難性が証明されているラベル選択付最小全域木問題 [4, 5] をこの問題に還元することで得られる．

定理 1 ラベル選択付最小連結全域部分グラフ問題は *NP* 困難である．

証明 $G=(V,E)$ をグラフ，w を繋ぐ場合の重み関数とする．繋がない場合の重み関数 $\overline{w}$ は，$\{v_1,v_2\}\in E$ ならば，すべての $l_1,l_2\in\Sigma$ について $\overline{w}(v_1,v_2,l_1,l_2)=0$，$\{v_1,v_2\}\in E$ ならば，すべての $l_1,l_2\in\Sigma$ について $\overline{w}(v_1,v_2,l_1,l_2)$ は未定義となるように定義する．連結全域部分グラフ $G'=(V,E')$ とラベル割り当て σ を，すべての連結全域部分グラフと G のラベル割り当ての組み合わせ中で，G' の重さが最小となるものとする．

$\overline{w}$ の定義より，もし，G' にサイクルが存在したとすると，サイクル中の辺のどれか1つを取り除いてできる連結全域部分グラフの（ラベル割り当て σ に対する）重さは同じはずである．このとき，取り除かれる辺は，繋ぐ場合の重みも繋がない場合の重みも0でなくてならないことを注意する．もしそうでなければ，G' の重さが最小であるという仮定に矛盾する．よって，重さを変えることなくサイクル中の辺を取り除くことができるので，G' と同じ重さの全域木 T を得ることができる．

全域木 T は，$G=(V,E)$ をグラフ，w を重み関数とした時のラベル選択付最小全域木問題の解である．ゆえに，既にNP困難性が証明されているラベル選択付最小全域木問題を，ラベル選択付最小連結全域部分グラフ問題に還元することができた． □

4 線形時間アルゴリズム

本章では，tree-widthが高々2のグラフに対する線形時間アルゴリズムを紹介する．

任意のtree-widthが高々2のグラフは，次に示す変形規則によって単一頂点のグラフに縮約できることが知られている [7]．（図2も見よ．）

1. グラフに頂点 v_2 が存在し，正確に1つの辺 $\{v_1,v_2\}$ に繋がっているとき，頂点 v_2 と辺 $\{v_1,v_2\}$ を取り除く．
2. グラフに頂点 v_2 が存在し，正確に2つの辺 $\{v_1,v_2\},\{v_2,v_3\}$ に繋がっているとき，頂点 v_2 と辺 $\{v_1,v_2\},\{v_2,v_3\}$ を取り除き，新しい辺 $\{v_1,v_3\}$ を加える．
3. 規則2によって，頂点 v_1,v_2 の間に多重辺ができてしまったら，多重辺を取り除き，新しい辺 $\{v_1,v_3\}$ を加える．

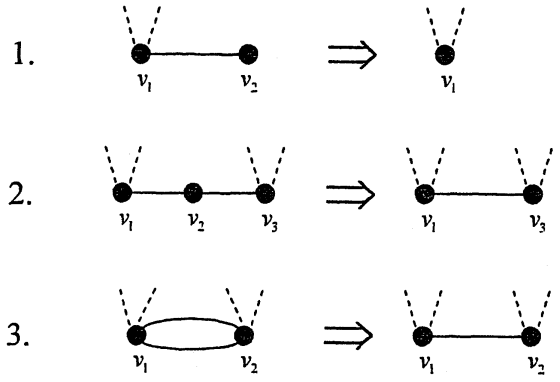

図 2: 変形規則.

Σをラベル記号の集合とする．以下のアルゴリズムは，グラフ $G=(V,E)$，Gの重み関数w，$\overline{w}$を受け取り，Gの tree-width が 2 以下であれば，連結全域部分グラフの最小の重さを返し，tree-width が 3 以上であれば，reject を返す.

Function 線形時間アルゴリズム

Input: グラフ $G=(V,E)$，Gの重み関数w，$\overline{w}$

Output: 連結全域部分グラフの最小の重さ，または，reject

for all $v \in V$ **do**
  for all $a \in \Sigma$ **do**
    $w'(v,a) := 0$
  end for
end for
while $|V| > 1$ **do**
  次数が 1 または 2 の頂点 $v_2 \in V$ を見つける
  if v_2 の次数が 1 で，1 つの辺 $\{v_1, v_2\}$ に繋がっている **then**
    $V := V - \{v_2\}$
    $E := E - \{\{v_1, v_2\}\}$
    for all $a \in \Sigma$ **do**
      $min := \infty$
      for all $b \in \Sigma$ **do**
        if $min > w(v_1, v_2, a, b) + w'(v_2, b)$ **then**
          $min := w(v_1, v_2, a, b) + w'(v_2, b)$
        end if
      end for
      $w'(v_1, a) := min$
    end for
  else if v_2 の次数が 2 で，2 つの辺 $\{v_1, v_2\}, \{v_2, v_3\}$ に繋がっている **then**
    $V := V - \{v_2\}$
    $E := E' - \{\{v_1, v_2\}, \{v_2, v_3\}\}$
    for all $a \in \Sigma$ **do**

```
24:       for all c ∈ Σ do
25:         min1 := ∞
26:         min2 := ∞
27:         for all b ∈ Σ do
28:           if min1 > w(v1, v2, a, b) + w'(v2, b) + w(v2, v3, b, c) then
29:             min1 := w(v1, v2, a, b) + w'(v2, b) + w(v2, v3, b, c)
30:           end if
31:           if min2 > w(v1, v2, a, b) + w'(v2, b) + w̄(v2, v3, b, c) then
32:             min2 := w(v1, v2, a, b) + w'(v2, b) + w̄(v2, v3, b, c)
33:           end if
34:           if min2 > w̄(v1, v2, a, b) + w'(v2, b) + w(v2, v3, b, c) then
35:             min2 := w̄(v1, v2, a, b) + w'(v2, b) + w(v2, v3, b, c)
36:           end if
37:         end for
38:         if {v1, v3} ∈ E then
39:           minA := w(v1, v3, a, c) + min1
40:           minB := w(v1, v3, a, c) + min2
41:           minC := w̄(v1, v3, a, c) + min1
42:           w(v1, v3, a, c) := min(minA, minB, minC)
43:           w(v3, v1, c, a) := w(v1, v3, a, c)
44:           w̄(v1, v3, a, c) := w̄(v1, v3, a, c) + min2
45:           w̄(v3, v1, c, a) := w̄(v1, v3, a, c)
46:         else
47:           E := E ∪ {{v1, v3}}
48:           w(v1, v3, a, c) := min1
49:           w(v3, v1, c, a) := w(v1, v3, a, c)
50:           w̄(v1, v3, a, c) := min2
51:           w̄(v3, v1, c, a) := w̄(v1, v3, a, c)
52:         end if
53:       end for
54:     end for
55:   else if 条件を満たす v2 ∈ V が存在しなかった then
56:     return reject
57:   end if
58: end while
59: v ∈ V とする
60: min := ∞
61: for all a ∈ Σ do
62:   if min > w'(v, a) then
63:     min := w'(v, a)
64:   end if
```

65: **end for**
66: **return** *min*

5 まとめと今後の課題

化学構造式OCRの認識精度向上に応用するため，ラベル選択付最小連結全域部分グラフ問題を考案した．ラベル選択付最小連結全域部分グラフ問題は，NP困難であることを示した．しかし，入力のグラフのtree-widthが2以下ならば，線形時間で動作する非常にシンプルなアルゴリズムの提案を行った．

現在，提案アルゴリズムを実装し，化学構造式OCRを開発中である．近日，プロジェクトのホームページ[8]で公開する予定である．

今後は，化学構造式生成文法を開発し，文法的特長量なども認識精度向上に利用することを検討する．

参考文献

[1] 特許電子図書館. `http://www.ipdl.inpit.go.jp/`.

[2] Akio Fujiyoshi, Koji Nakagawa, and Masakazu Suzuki. Robust method of segmentation and recognition of chemical structure images in cheminfty. In *Pre-Proceedings of the 9th IAPR International Workshop on Graphics Recognition (GREC 2011)*, 2011.

[3] Daniel Karzel, Koji Nakagawa, Akio Fujiyoshi, and Masakazu Suzuki. Inconsistency-driven chemical graph construction in cheminfty. In *Pre-Proceedings of the 9th IAPR International Workshop on Graphics Recognition (GREC 2011)*, 2011.

[4] 藤芳明生, 鈴木昌和. ラベル選択を有する最小全域木問題. 2009年度冬のLAシンポジウム.

[5] Akio Fujiyoshi and Masakazu Suzuki. Minimum spanning tree problem with label selection. *IEICE Trans. Inf. & Syst.*, Vol. E94-D, No. 2, pp. 233–239, 2011.

[6] ChEMBL. `https://www.ebi.ac.uk/chembl/`.

[7] Stefan Arnborg and Andrzej Proskurowski. Characterization and recognition of partial 3-trees. *SIAM J. Algebraic Discrete Methods*, Vol. 7, No. 2, pp. 305–314, 1986.

[8] ChemInfty. `http://www.inftyproject.org/jp/ChemInfty/`.